Supplementary Online Material

Long-term balancing selection in *LAD1* maintains a missense trans-species polymorphism in humans, chimpanzees and bonobos

João C. Teixeira[1]*, Cesare de Filippo[1]*, Antje Weihmann[1], Juan R. Meneu[1], Fernando Racimo[2], Michael Dannemann[1], Birgit Nickel[1], Anne Fischer[3], Michel Halbwax[4], Claudine Andre[5], Rebeca Atencia[6], Matthias Meyer[1], Genís Parra[1], Svante Pääbo[1] and Aida M. Andrés[1]

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany

[2]Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA

[3]International Center for Insect Physiology and Ecology, Nairobi 30772-00100, Kenya

[4]Clinique vétérinaire du Dr. Jacquemin, 94700 Maisons-Alfort, France

[5]Lola Ya Bonobo sanctuary, Kinshasa, Democratic Republic Congo

[6]Réserve Naturelle Sanctuaire à Chimpanzés de Tchimpounga, Jane Goodall Institute, Pointe-Noire, Republic of Congo

*Authors contributed equally

Corresponding author: Aida M. Andrés (aida_andres@eva.mpg.de)

I – Identification, filtering and validation of shSNPs

After performing genotype calling with GATK (McKenna et al. 2010), we proceeded to filter putative false positive variants (see Materials and Methods). We initially identified shSNPs as orthologous SNPs that showed the same two alleles in all three species. This set did not include orthologous polymorphic positions for which at least two species showed different alternative alleles, which we instead define as coincident SNPs (cSNPs) as the position is polymorphic across these species but the alleles are different (see section III).

Because shared SNPs (shSNPs) across different species might be enriched for systematic sequencing errors, we adopted additional filtering criteria only on shSNPs (see Materials and Methods) to exclude such errors. After these extra filtering, we uncovered a total of 138 coding shSNPs that are present in the three species, and that are not in CpG sites. Due to potential problems arising from incorrectly mapped reads, we performed additional filtering on the shSNPs and excluded shSNPs: 1) in Hardy-Weinberg disequilibrium ($p<0.05$); 2) that fall in regions with unusually high coverage (5% upper tail of the coverage distribution of all SNPs); and 3) that are located in regions with 24mer mappability lower than 100% (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability). Although these hard cutoffs potentially resulted in the removal of some true positives, they should largely remove false positives (see section on Sanger sequencing validation below). We obtained a total of 13 shSNPs in all three species, 9 of which (69.23%) result in a non-synonymous change and alter protein sequence. 10 shSNPs (76.92%) are located in three *HLA* genes that belong to the MHC region on chromosome 6, which is the best-established example of balancing selection in vertebrates (Klein et al. 1993; Graser et al. 1996; Asthana, Schmidt, Sunyaev 2005; Loisel et al. 2006; Cutrera, Lacey 2007; Kikkawa et al. 2009; Leffler et al. 2013; Sutton et al. 2013). The remaining three shSNPs are located in two genes, *LAD1* (*ladinin 1*) and *TNFRSF10D* (*tumor necrosis factor receptor superfamily, member 10d*), in chromosomes 1 and 8, respectively.

Besides being the consequence of balancing selection, shSNPs can also occur due to species proximity. Here, we can rule out the effects of a recent split or incomplete lineage sorting as being causative for the presence of shSNPs among the three species because the vast majority of neutrally evolving polymorphisms in the ancestral species of humans, chimpanzees and bonobos are expected to have drifted to fixation (see section II) (Clark 1997).

Finally, a shSNP may also be the result of recurrent mutation in the different lineages. Because a true trSNP must fall in a region where sequences cluster by allele rather than by species, we only considered shSNPs that cluster in an allelic tree and not in a species tree (see Materials and Methods and Results) as candidate trans-species SNPs (trSNPs). Only three trSNPs, all located in *HLA-DQA1*, show surrounding regions that do not cluster in an allelic tree (p=0.90). Finally, and to confirm that these are true SNPs rather than genotype errors, we further validated the candidate trSNPs outside the HLA genes using Sanger sequencing.

Therefore, we identified 10 trSNPs, seven of them in the HLA loci, one in the gene *LAD1*, and 2 in the gene *TNFRSF10D*. Because the HLA is a known target of selection we focus below on the three non-HLA trSNPs.

Sanger sequencing validation

We produced Sanger resequencing data for regions surrounding the 3 candidate trSNPs in all three species. A total of 18 bonobos, 19 chimpanzees and 18 humans were used in this analysis. The primers were designed specifically for each species by taking into account the substitutions identified in our dataset. All primer pairs were designed using Primer3 (Rozen, Skaletsky 2000), ensuring a single amplification product for the majority of the fragments (amplicon sizes vary from 504bp to 642bp). Additional sets of primers and different primer combinations were used in cases where a PCR reaction failed or multiple bands prevented effective sequencing. PCR reactions were performed using Herclase II Fusion (Agilent Technologies), and following manufacturer's recommendations. After amplification, PCR products were purified using SPRI beads. Sequencing reactions were carried using the BigDye terminator v1.1 Cycle Sequencing chemistry (Applied Biosystems), and purified by ethanol/sodium acetate precipitation. Sanger sequencing was performed using an ABI 3730 sequencer (Applied Biosystems). All sequences were analyzed using the Sequencing Analysis software provided with the instrument (Applied Biosystems). We were able to validate 2 out of the 3 candidate trSNPs, one in *LAD1* (chr1: 201355761) and the other in *TNFRSF10D* (chr8: 23003292) but unable to validate the other candidate trSNP located in *TNFRSF10D* (chr8: 22995487).

We also attempted to validate some additional shSNPs that did not pass the Hardy-Weinberg equilibrium and mappability filters. Human showed the highest percentage of validated shSNPs (36.21%) and the lowest percentage of defined false positives (34.48%), with 29.31% of shSNPs that we could not ascertain with Sanger, and remained ambiguous. 10.35% of shSNPs were validated in bonobo, the same as in chimpanzee; moreover, 60.34% of the shSNPs could not be validated and 29.31% could not be ascertained in bonobo, and in chimpanzee 46.55% were not validated 43.10% were not ascertained. Sanger validation was hampered by two main problems: first, the difficulty to obtain clean bands of expected size by PCR in some of the species; second, the presence of short non-annotated segmental duplications in some species. Specifically, in about 50% of the shSNPs that failed validation we observed repeated regions of variable length (20-50bp) around the SNP that make these shSNPs difficult to validate by Sanger sequencing. This is not surprising as these SNPs did not pass all our quality and uniqueness filters above.

BLAT analysis

We performed a BLAT analysis as a final step in order to ensure these two trSNPs were real using UCSC's 'BLAT Search Genome' tool. We used the -/+25bp surrounding each of the two shSNPs as query, and performed a BLAT of each sequence against the human genome (hg19), first using the reference allele and then using the alternative allele for each shSNP. After this, we repeated the analysis using the chimpanzee genome (PanTro4). We found that, for the shSNP located in *LAD1*, the surrounding sequence maps uniquely (for both alleles) both in the human and the chimpanzee genomes so we could confirm this is indeed a trans-species polymorphism. As for the SNP located in *TNFRSF10D*, we found that the sequence containing the reference or the alternative alleles maps uniquely to hg19, but is likely duplicated in PanTro4. We obtain two matches with the same sequence identity that are located ~50kb upstream the predicted position of the SNP. Although this does not discard this position as a SNP in the other species (or in chimpanzee) we conservatively removed it from further analyses.

II – Trans-species Polymorphism due to Neutral Identity by Descent

In a coalescent genealogy, the probability of a neutral polymorphism shared between humans and chimpanzees due to identity by descent is very low (Leffler et al. 2013). This probability depends on at least two human lineages and two chimpanzee lineages not coalescing before the human-chimpanzee split time, and on a particular order of coalescence events in the ancestral population along with the occurrence of a mutation on the correct part of the genealogy (Wiuf et al. 2004; Segurel et al. 2012). We targeted SNPs evolving under long-term balancing selection by focusing on trans-species polymorphisms shared between humans, chimpanzees and bonobos, so it is necessary to calculate the probability that such polymorphisms are neutral.

For this, we begin by studying the properties of a trans-species polymorphism in a coalescent genealogy of 2 lineages per species. First, looking backwards in time, for a trSNP to occur, a general requirement is that none of the pairs of lineages of each species coalesce during their species-specific history. Assuming this is the case, there are three different types of scenarios that could result in a polymorphism shared between the three species: a) none of the two chimpanzee and two bonobo lineages coalesce from the time of the chimpanzee-bonobo split to the time of the human-chimpanzee-bonobo split; b) a single chimpanzee lineage coalesces with a single bonobo lineage during this time; and c) a single chimpanzee lineage coalesces with a single bonobo lineage, and a different chimpanzee lineage coalesces with a different bonobo lineage during this time. These different scenarios are shown in figure S1. Moreover, it is then necessary for a mutation to occur in the correct lineage in the ancestral population, such that it leads to a pattern consistent with a trans-species polymorphism. The probability of occurrence of such a mutation varies according to the number of lineages reaching the human-chimpanzee-bonobo ancestral population. In other words, different tree topologies might have different probabilities of producing a neutral trans-species polymorphism.

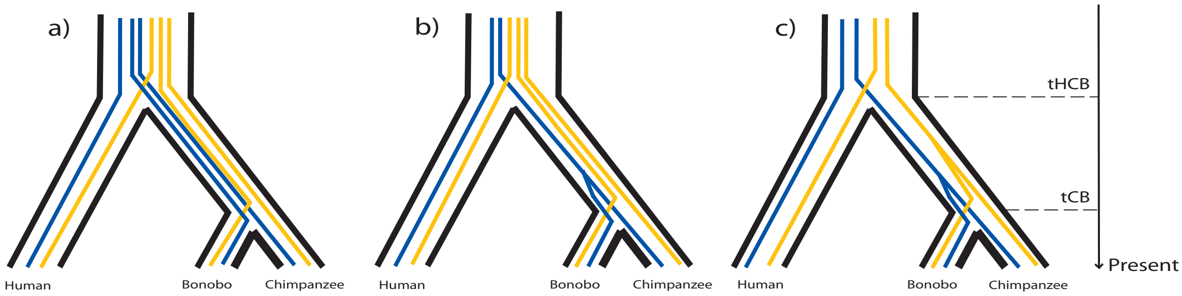


Figure S1 – Genealogical scenarios allowing for a trans-species polymorphism shared between humans, chimpanzees and bonobos. In all three examples, the mutation must arise in the ancestral population before the human-chimpanzee-bonobo split time. The blue and yellow coloring of different lineages is arbitrary, and does not denote derived or ancestral states. tHCB and tCB represent the split times of human-chimpanzee-bonobo and chimpanzee-bonobo, respectively.

Below, we estimate the probability that a human polymorphism is also a trans-species polymorphism in chimpanzees and bonobos, under neutrality. We assume that population sizes stay constant within each species (but not necessarily across species), that there is no population structure within species or migration among species, that there is no recurrent mutation, and that the number of sampled chromosomes is small relative to the whole population for each species. The probability we obtain will not depend on the value of the mutation rate per site (so long as the mutation rate is constant along the genealogy, which we assume for simplicity), as it will be a ratio of two terms which both contain the mutation rate, and so the rate will cancel out. We first define the following terms:

Term	**Definition**
N_A	Human-chimpanzee-bonobo ancestral population size
N_H	Human population size since the population split with chimpanzees and bonobos
N_C	Chimpanzee population size since the chimpanzee-bonobo population split time
N_B	Bonobo population size since the chimpanzee-bonobo population split time
N_{CB}	Population size of chimpanzees and bonobos after the split with humans but before the split between each other
t_{HCB}	Population split time (in generations) of humans and chimpanzees+bonobos
t_{CB}	Population split time (in generations) of chimpanzees and bonobos
t_X	$(t_{HCB} - t_{CB})/(2*N_{CB})$

$P_{Htsp}(x,x+\Delta x)$	Probability of finding a site where 2 human chromosomes are different, 2 bonobo chromosomes are different and 2 chimpanzee chromosomes are different in a region of length Δ x
$P_{Hhum}(x,x+\Delta x)$	Probability of finding a site where 2 human chromosomes are different in a region of length Δ x
$P_{Ptsp}(x,x+\Delta x)$	Probability of finding a site where humans are polymorphic, bonobos are polymorphic and chimpanzees are polymorphic in a region of length Δ x
$P_{Phum}(x,x+\Delta x)$	Probability of finding a site where humans are polymorphic in a region of length Δ x
P_{TSPHET}	Probability that 2 bonobo chromosomes are different and 2 chimpanzee chromosomes are different at a site, given that 2 human chromosomes are different at that site
P_{FINAL}	Probability that a site is polymorphic in both bonobos and chimpanzees, given that it is polymorphic in humans
u	Mutation rate per site per generation
g(n,j,t)	Ancestral process of the coalescent: probability of there being j lineages at time t in the past, given that there were n lineages at time 0, measuring time in coalescent units (Tavare 1984)

We define ETA[k] to be the expectation for the inter-coalescence time (in generations) while there are k lineages in the human-chimpanzee-bonobo ancestral population:

$$\mathrm{ETA}[k] = \frac{2N_A}{\binom{k}{2}}$$

We also define ETH[2, N_H, N_A, t_{HCB}] to be the expectation for the time until coalescence (in generations) of 2 lineages sampled in the human population in the present (Griffiths, Tavare 1994):

$$\mathrm{ETH}[2, N_H, N_A, t_{HCB}] = 2N_H \int_0^{\infty} e^{-\int_0^{v} f(z,N_H,N_A,t_{HCB})dz} dv$$

where f(z, N_H, N_A, t_{HCB}) is a piecewise constant function defined as 1 when z <= $t_{HCB}/(2N_H)$ and N_H/N_A when z > $t_{HCB}/(2N_H)$.

We begin by obtaining the probability that a site that is heterozygous in 2 human chromosomes is also heterozygous in 2 bonobo chromosomes and in 2 chimpanzee chromosomes:

$$P_{TSPHET} = \lim_{\Delta x \to 0} \frac{P_{Htsp}(x, x+\Delta x)}{P_{Hhum}(x, x+\Delta x)} = \frac{(e^{-t_{HCB}/(2N_H)})(e^{-t_{CB}/(2N_C)})(e^{-t_{CB}/(2N_B)})PX}{u*2*ETH[2, N_H, N_{A,} t_{HCB}]}$$

where $PX = [g(4,4,t_X)*PA + g(4,3,t_X)*(2/3)*PB + g(4,2,t_X)*(2/7)*PC]$

$$PC = \frac{2u*PO}{9}$$

$$PB = \frac{u}{10}\left[ETA[4] + PO\right]$$

$$PA = \frac{4}{5}PB$$

and

$$PO = (3*ETA[3] + 2*ETA[2])$$

Here, PA, PB and PC correspond to the probabilities of a mutation occurring in the correct lineage in the human-chimpanzee-bonobo ancestral population, given scenarios a), b) and c), respectively. The ancestral process functions g(n,j,t) in each term of PX allow us to calculate the probability of each scenario.

Following Leffler et al. (2013), we can approximate the probability that a site that is polymorphic in a human sample is also polymorphic in a sample of bonobos and a sample of chimpanzees in the following way:

$$P_{FINAL} = \lim_{\Delta x \to 0} \frac{P_{Ptsp}(x, x+\Delta x)}{P_{Phum}(x, x+\Delta x)} \approx \frac{(e^{-(t_{HCB}-2N_H)/(2N_H)})(e^{-(t_{CB}-2N_C)/(2N_C)})(e^{-(t_{CB}-2N_B)/(2N_B)})PX}{u * 2 * ETH'[N_H, N_{A,} t_{HCB}]}$$

where $ETH'[N_H, N_A, t_{HCB}] = 2N_H \int_0^\infty e^{-(\int_0^v f(z, N_H, N_A, t_{HCB})dz)+1} dv$

We fixed the relevant population size and split time parameters at the values estimated in (Prado-Martinez et al. 2013): $N_A = 55{,}000$, $N_H = 8{,}000$, $N_C = 30{,}000$, $N_B = 5{,}000$, $N_{CB} = 30{,}000$, $t_{HCB} = 250{,}000$, $t_{CB} = 40{,}000$.

Using these values, we obtain that P_{FINAL} is equal to $4.05\text{x}10^{-10}$. We can compare this probability to the probability of seeing a polymorphism in a sample of chimpanzees, given that the site is polymorphic in a sample of humans, as in Leffler et al. (2013). Let us denote this probability as P_{HC}:

$$P_{HC} \approx \frac{(e^{-(t_{HCB}-2N_H)/(2N_H)})(e^{-(t_{CB}-2N_C)/(2N_C)})(e^{-(t_{HCB}-t_{CB})/(2N_{CB})})PC}{u * 2 * ETH'[N_H, N_A, t_{HCB}]}$$

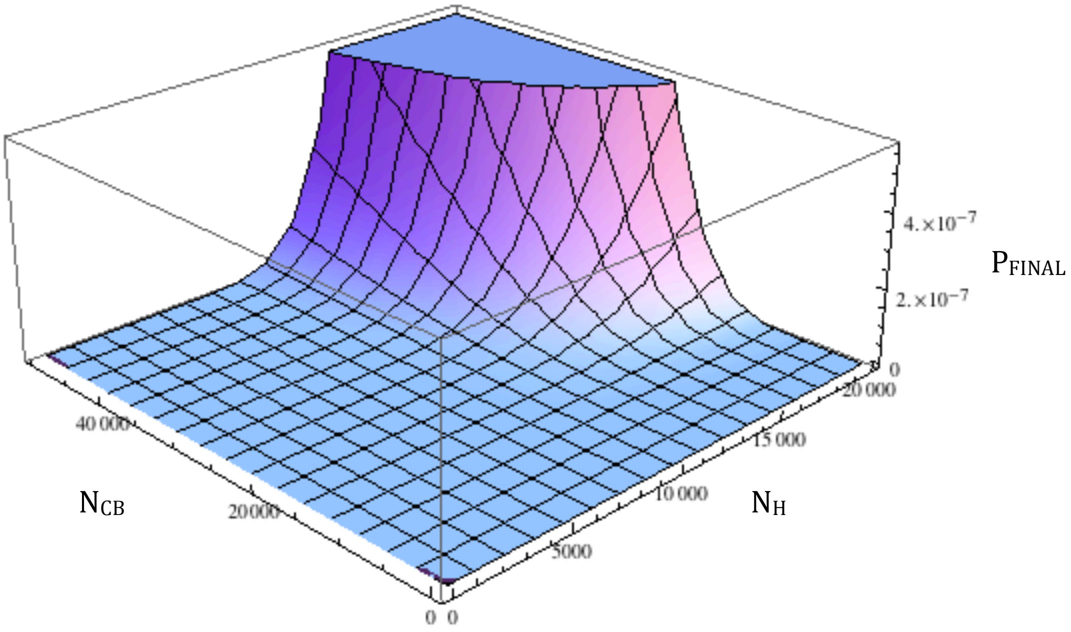
Using the same fixed parameters as above, we obtain that this probability equals $1.58\text{x}10^{-8}$, which is 39 times larger than P_{FINAL}.

Additionally, using an analogous calculation to P_{HC} and assuming $N_A = N_{CB} = 30{,}000$, we obtained the probability that a site is polymorphic in bonobos given that it is polymorphic in chimpanzees (= 0.0085) as well as the probability that a site is polymorphic in chimpanzees given that it is polymorphic in bonobos (= 0.046). The probability is higher in the latter case because $N_B < N_C$. This implies that the denominator in the first case is larger than in the second case, while the numerator stays the same.

We can also vary some of these parameters and observe the behavior of P_{FINAL} under different input values. For example, we plotted P_{FINAL} in Figure S1 as a function of the human population size (ranging from 0 to 20,000) and the chimpanzee-bonobo population size during their shared history (ranging from 0 to 200,000). As expected, as population sizes increase (making recent coalescences less likely) this probability also increases. Interestingly, the $\log_{10}$ of this probability drops sharply when either of the two population sizes are small (~1,000), because coalescent events tend to happen very early in populations of those sizes, and so trans-species polymorphisms become extremely unlikely.

In Figure S2, we show P_{FINAL} as a function of the human-chimpanzee-bonobo split time (t_{HCB}, ranging from 100,000 to 300,000 generations) and the chimpanzee-bonobo split time (t_{CB}, ranging from 0 to 100,000 generations). Again, as expected, this probability decreases as a function of the split times.

According to our approximation, the probability of a segregating site in humans being a trans-species polymorphism with chimpanzee and bonobo is $4.05\text{x}10^{-10}$. Given that we find 121,904 SNPs in humans, we expect 0.00005 (basically none) shSNPs with chimpanzee and bonobo under neutrality. Hence, these results suggest that trSNPs are unlikely to arise by neutrality, and other forces, like long-term balancing selection, must be responsible for their maintenance.



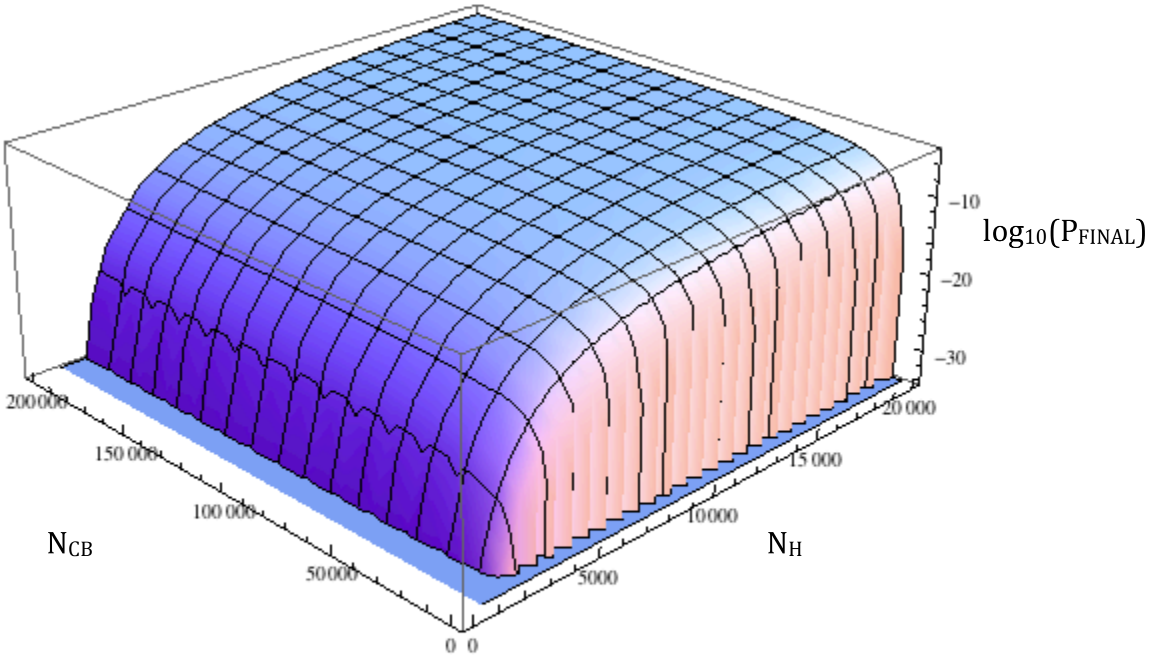


Figure S1. Top panel: P_{FINAL} plotted as a function of the human population size (ranging from 0 to 20,000) and the chimpanzee-bonobo population size during their shared history (ranging from 0 to 200,000). Bottom panel: $\log_{10}(P_{FINAL})$ plotted as a function of the same parameters. All other parameters are held fixed at the values estimated in Prado-Martinez et al. (2013).

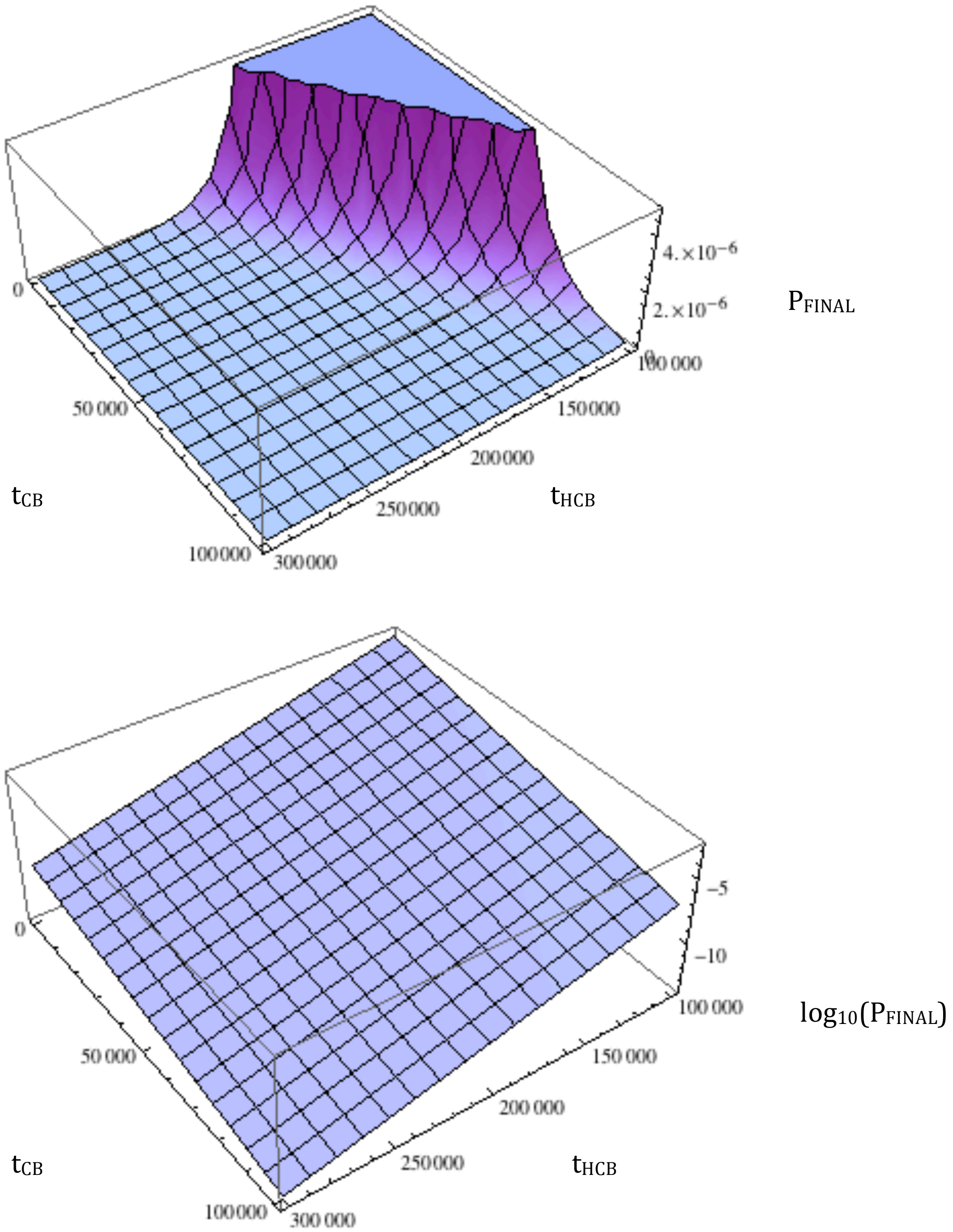


Figure S2. Top panel: P_{FINAL} plotted as a function of the human-chimpanzee-bonobo split time (t_{HCB}, ranging from 100,000 to 300,000 generations) and the chimpanzee-bonobo split time (t_{CB}, ranging from 0 to 100,000 generations). Bottom panel: $\log_{10}(P_{FINAL})$ plotted as a function of the same parameters. All other parameters are held fixed at the values estimated in Prado-Martinez et al. (2013).

We also simulated 10 sets of 10,000 genealogies for different demographic scenarios in ms (Hudson 2002) to verify our analytical expression for P_{TSPHET} was correct (Figure S3). For each set, we obtained the average P_{TSPHET} across genealogies. In the different scenarios, we used shorter population split times than in the human-chimpanzee-bonobo scenario due to the computational cost of obtaining a branch where a trans-species polymorphism can appear when population split times are far in the past. The simulated and analytical values differ most when the true value is small (e.g. Scenario C), because in those cases most of the sampled simulated genealogies contain no branches where a trans-species polymorphism is possible and so sparse sampling of the correct genealogies increases the error in the simulation estimates.

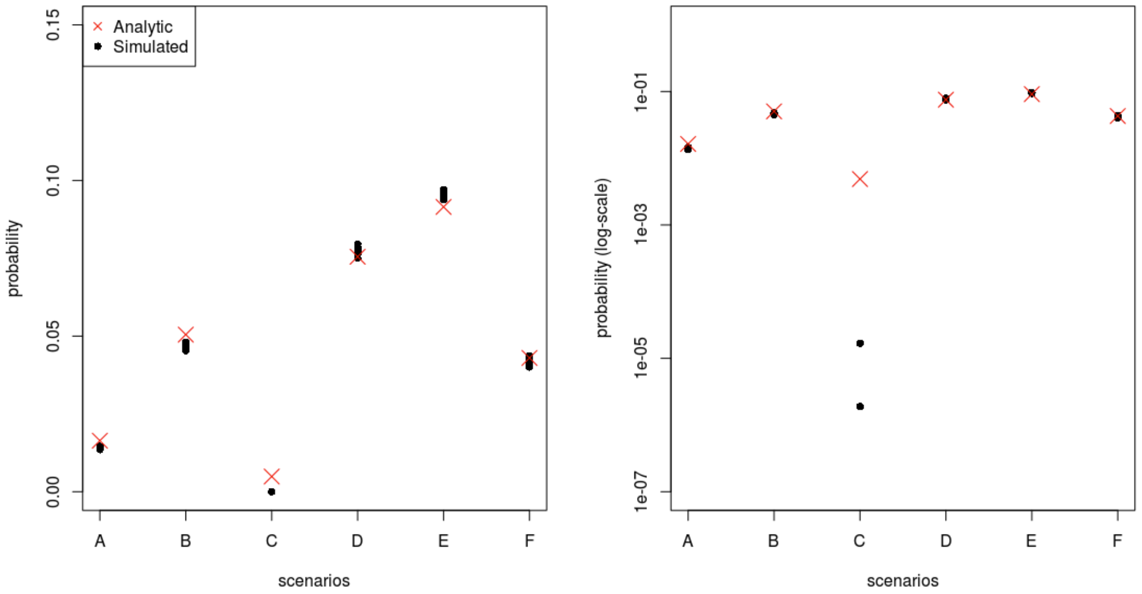


Figure S3. Analytic and simulated values for P_{TSPHET} under different demographic scenarios. The simulated values were obtained from the average of a set of 10,000 simulated genealogies, and we plotted 10 sets per scenario. The right panel shows the same values as the left panel but with a log-scaled probability on the y-axis. The parameters used for each simulated scenario were as follows: A) $N_H = N_C = N_B = N_{CB} = N_A = 10{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. B) $N_H = N_C = N_B = N_{CB} = N_A = 10{,}000$; $t_{HCB} = 10{,}000$; $t_{CB} = 1{,}000$. C) $N_H = N_C = N_B = N_{CB} = N_A = 10{,}000$; $t_{HCB} = 30{,}000$; $t_{CB} = 10{,}000$. D) $N_H = 10{,}000$; $N_C = N_B = N_{CB} = N_A = 50{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. E) $N_H = 10{,}000$; $N_C = N_B = 50{,}000$; $N_{CB} = N_A = 100{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. F) $N_H = 8{,}000$; $N_{CB} = N_C = 30{,}000$; $N_B = 5{,}000$; $N_A = 55{,}000$; $t_{HCB} = 20{,}000$; $t_{CB} = 5{,}000$. In all but two of the sets of Scenario C, all trees simulated under this scenario contained no branches where a trans-species polymorphism is possible and so sparse sampling of simulations leads to underestimation of the true value for P_{TSPHET}. The other 8 sets therefore had an average simulated $P_{TSPHET} = 0$. The right-hand plot only shows values of average P_{TSPHET} for the two sets in Scenario C where at least one tree contained a trans-species polymorphism (average simulated $P_{TSPHET} > 0$).

III – Evidence of an excess of coincident SNPs

One of the caveats of using trans-specific polymorphisms to target signatures of long-standing balancing selection is recurrent mutation, which may result in shSNPs. Therefore, when identifying trans-species polymorphisms, it is key to remove positions that have mutated independently in the different species. It has long been known that sites in the human genome show differences in mutation rates (Benzer 1961). In mammals, the best-characterized example of increased mutation rate is known as the 'CpG effect', in which the mutation rate of CG dinucleotides is increased by 10-fold compared to other sites (Hodgkinson, Eyre-Walker 2011). The cytosine in these dinucleotides is often methylated and, therefore, prone to suffer deamination to thymine, resulting in a C -> T transition (and G->A on the complementary DNA-strand) (Coulondre et al. 1978).

Mutation rate heterogeneity is not limited to CpG sites. Hodgkinson et al. (2009) showed that mutation rate biases increase the occurrence of SNPs at orthologous positions (coincident SNP – cSNP) in humans and chimpanzees. The authors found an excess of cSNPs in humans and chimpanzees, and suggested that sequence context around cSNPs are driving this pattern (Hodgkinson, Eyre-Walker 2010). The origin and exact nature of this sequence dependence remains cryptic. A similar analysis showed identical results when looking at orthologous substitutions in humans and chimpanzees using macaque as an out-group (Johnson, Hellmann 2011).

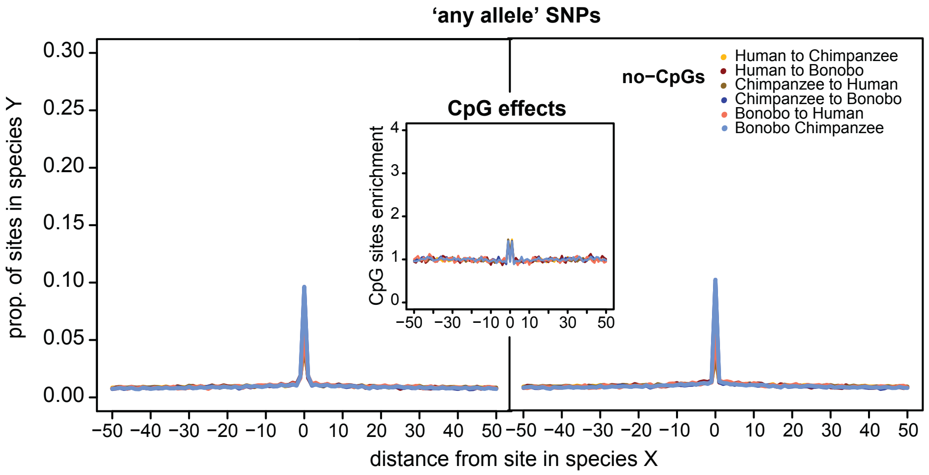
To understand the influence of recurrent mutation in our dataset we performed similar analyses. First, we identified every SNP found in the three species and divided them into:

- Coincident SNPs (cSNPs) – a SNP found in orthologous positions in at least two species, regardless of the alternative allele,
- Shared SNPs (shSNPs) – a SNP found in orthologous positions in at least two species and presenting the same alternative allele.

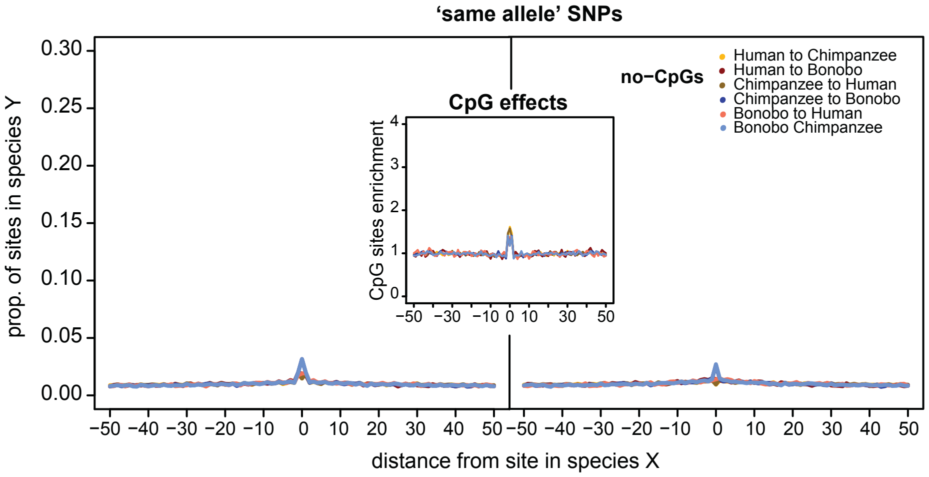
We also split each set of SNPs into CpG SNPs and non-CpG SNPs. Separate analyses were performed for coding, non-coding, synonymous and non-synonymous SNPs. Following Hodgkinson et al. (2009), for each SNP we analyzed the surrounding region (-50bp; +50bp) and identified additional SNPs in the interval, in each species. The results are presented in figure S4.

Similar to what had been previously reported, we find an excess of cSNPs relative to near-by regions for all pairwise comparisons. As expected, this pattern is stronger when looking at chimp-bonobo cSNPs due to the recent split between the two species that allows for the persistence of neutral polymorphisms in both species. After excluding SNPs in CG dinucleotides in order to account for CpG effects, we observe a decrease in the number of cSNPs for all pairwise comparisons. Predictably, the biggest effects are seen on adjacent positions to the cSNPs (-/+ 1bp). The pattern is noisier after dividing the data into synonymous and non-synonymous SNPs because of the lower number of sites, but we observe that, overall, controlling for CpG effects might not be enough to reduce recurrent mutation biases since the peaks at the central site (0bp) are still very high compared to surrounding regions. Nevertheless, this effect is much weaker when considering only shSNPs: for all comparisons, the excess of SNPs at orthologous positions almost disappears when we require the same alternative allele to be observed in both species (particularly in the human-chimpanzee and human-bonobo pairwise comparisons). Hence, these results show that while there is a neutral enrichment of cSNPs, the number of shSNPs that share the same two alleles in two species is barely higher than expected.

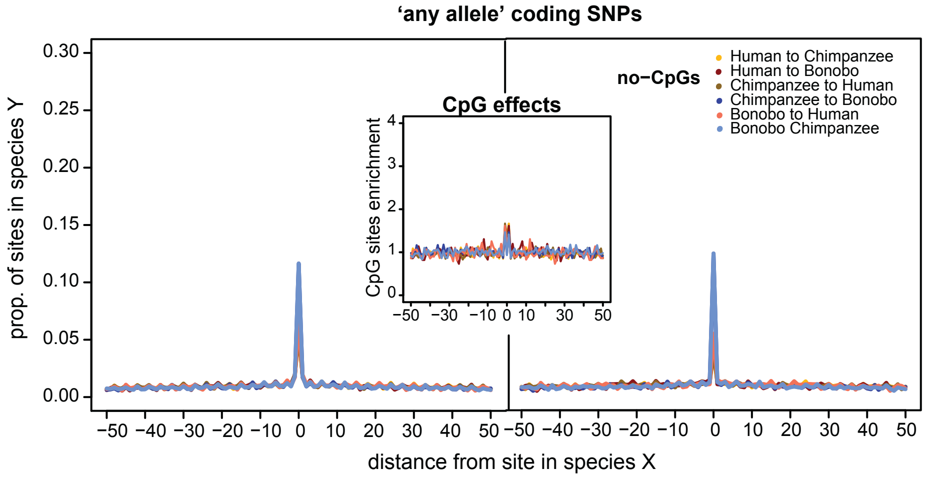
Only shSNPs can represent trans-species polymorphisms, so we focus on shSNPs alone, discard SNPs present in CG dinucleotides (potential CpG sites), and use additional signatures from patterns of linked variation to distinguish old polymorphisms from recurrent mutations.



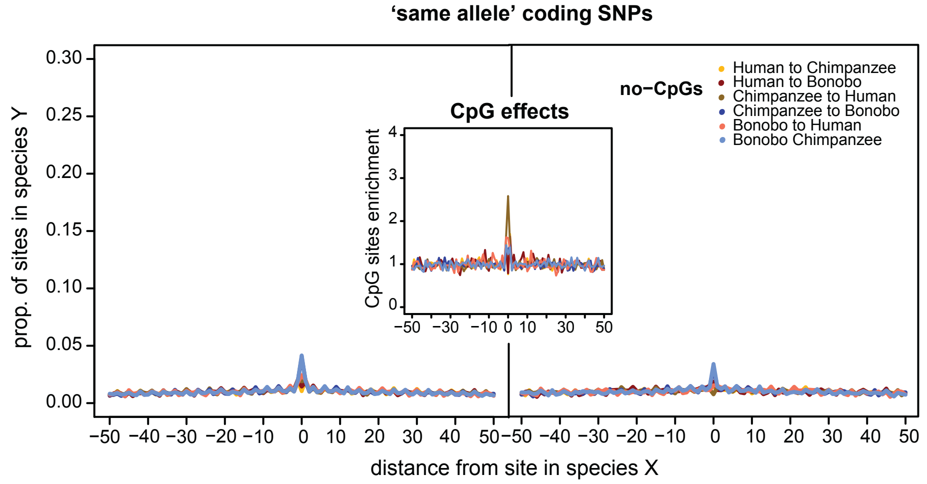
'any allele' SNPs



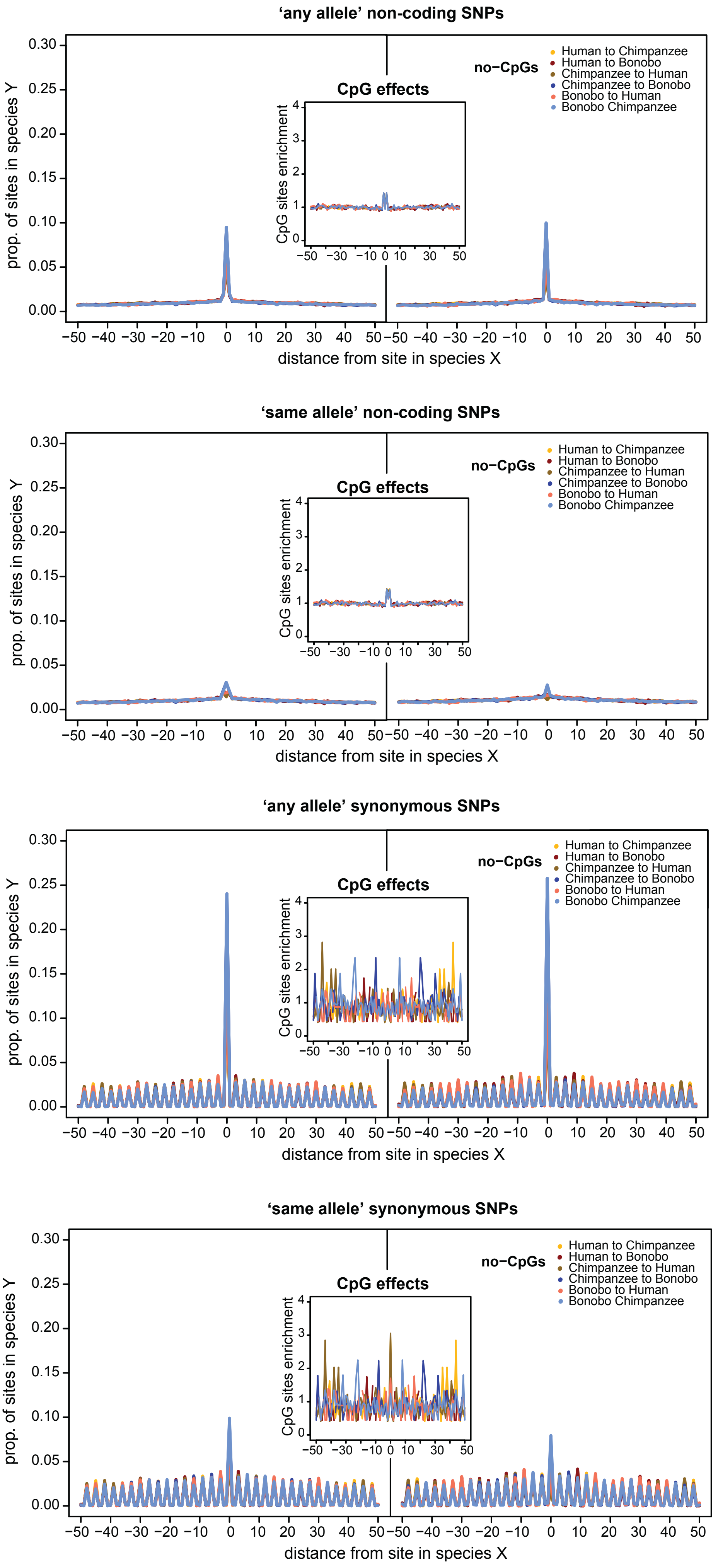
'same allele' SNPs



'any allele' coding SNPs



'same allele' coding SNPs

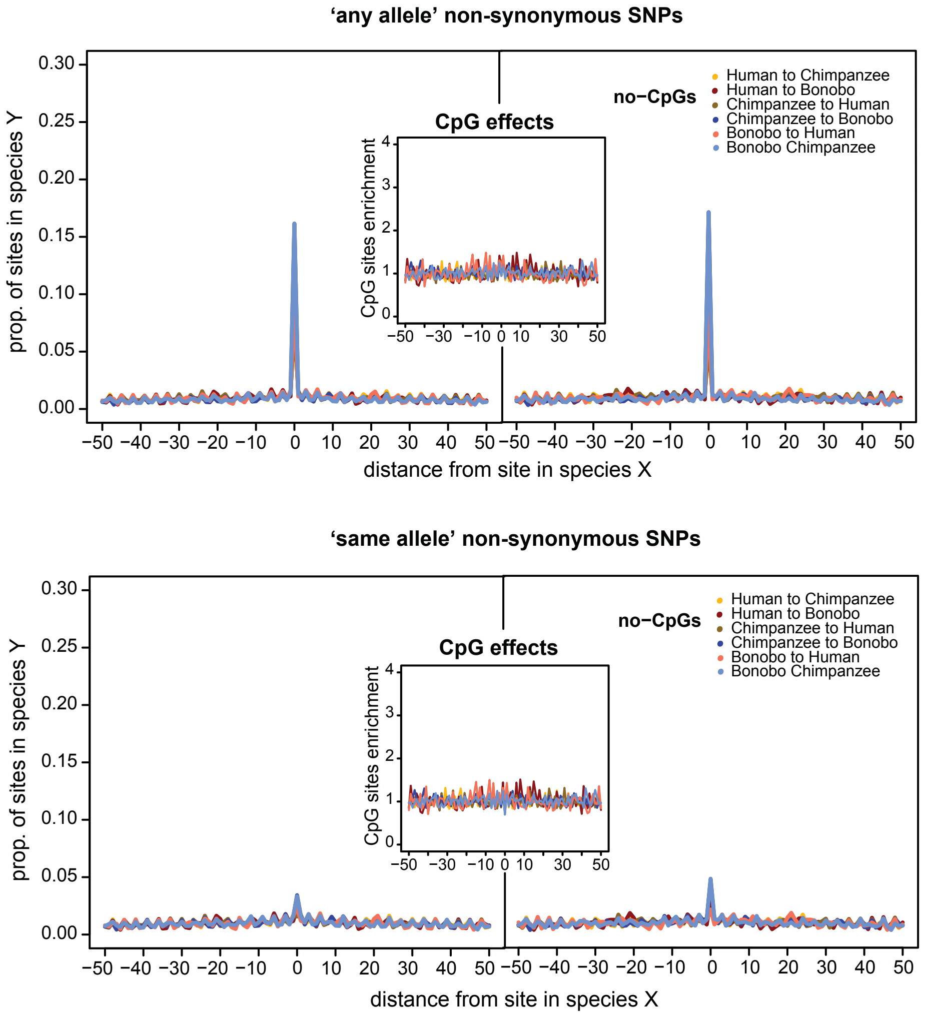


'any allele' non-coding SNPs

'same allele' non-coding SNPs

'any allele' synonymous SNPs

'same allele' synonymous SNPs



‘any allele’ non-synonymous SNPs

‘same allele’ non-synonymous SNPs

FigureS4 – Proportion of SNPs in one species according to the distance to a SNP in another species, shown for all six pairwise alignments. Analyses considering cSNPs and shSNPs are shown together with (on the left) and without (on the right) CpGs. The ‘CpG effect’ (ratio between CpG SNPs to non-CpG SNPs) is superimposed in the middle. Comparisons include a) all SNPs (coding+non-coding), b) coding SNPs, c) non-coding SNPs, d) synonymous SNPs, and e) non-synonymous SNPs. Human-to-Chimpanzee refers to human SNPs surrounding a SNP in chimpanzee and Chimpanzee-to-Human refers to chimpanzee SNPs surrounding a SNP in human, and likewise for the remaining four pairwise comparisons.

IV – Ratio of polymorphism to divergence in candidate genes

The patterns of diversity in a region surrounding a balanced polymorphism can be used to determine whether a given locus evolved under selection. In the particular case of long-standing balancing selection, the coalescent times of selected loci will be older than those of neutrally evolving ones, which, considering a constant mutation rate, results in an excess of polymorphism and deficiency of divergence linked to the selected variant (Charlesworth 2006). We calculated the polymorphism-to-divergence ratio *PtoD = p/(d+1)*, where *p* is the number of polymorphisms found in a species and *d* the number of fixed differences between species. This statistic allowed us to infer whether the set of candidate genes was significantly more polymorphic when compared to control genes (empirical genomic distribution) and, at the same time, control for heterogeneity in the mutation rates (since both SNPs and single-nucleotide substitutions – SNSs – are included).

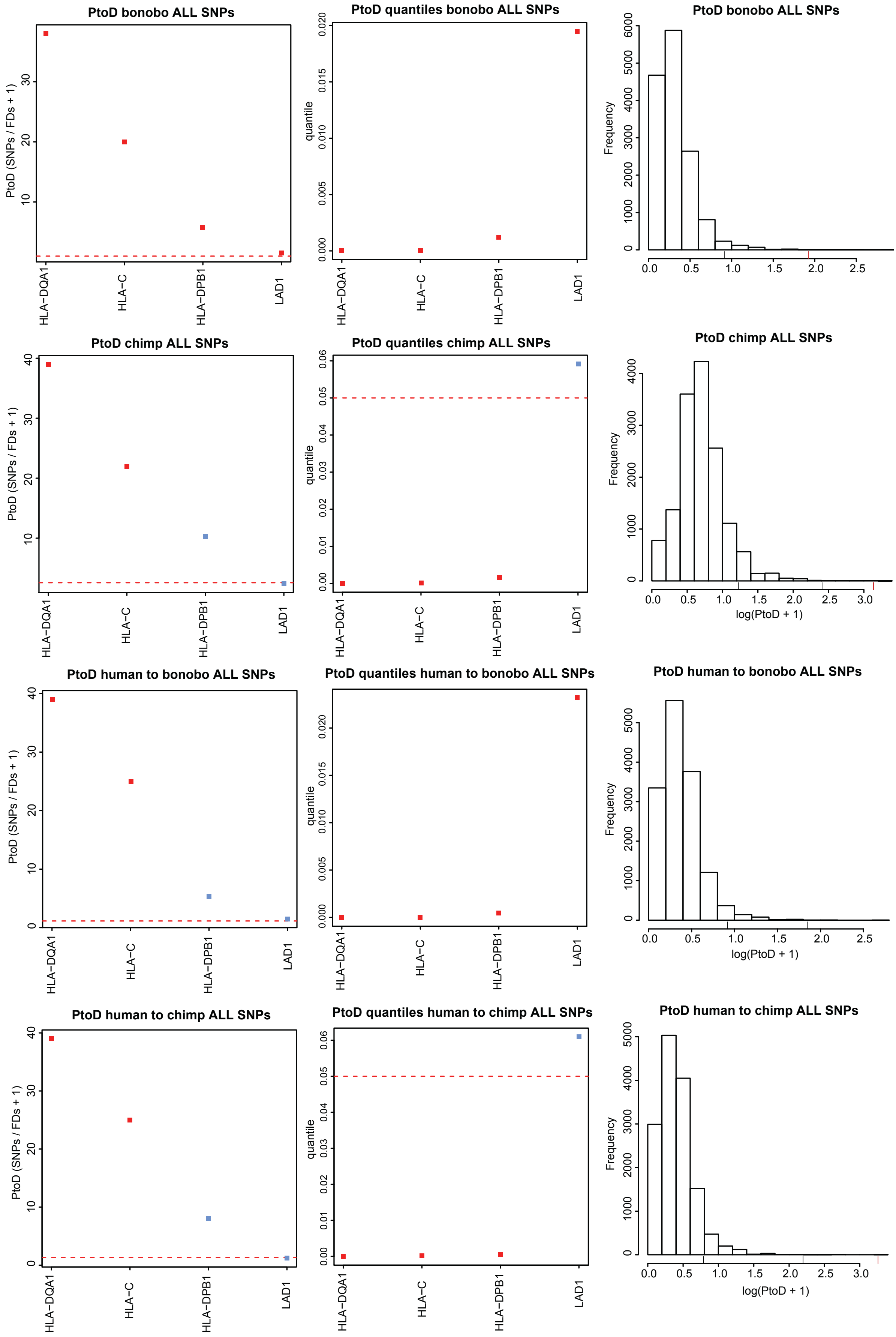
PtoD ratios were calculated for all genes considered as informative (i.e. all the genes that had at least one SNP or one substitution in our dataset after quality filtering). We then computed 2-tail Mann-Whitney U (MW-U) tests using R to assess whether the distribution of the average *PtoD* of candidate genes (*HLA-C*, *HLA-DQA1*, *HLA-DPB1* and *LAD1*) was significantly greater than the distribution of control genes, a classical signature of long-term balancing selection (Andres 2011). After comparing the *PtoD* values in the two groups, we sequentially removed the top candidate (i.e. one gene each time) from the candidate's group and recalculated MW-U p-values maintaining the control group unaltered. This approach allowed us not only to control for the potential effects of a few known highly diverse candidates (i.e. *HLA* genes), but at the same time to infer if the PtoD distributions remained significantly different after removing them. We calculated *PtoD* values using three different sets:

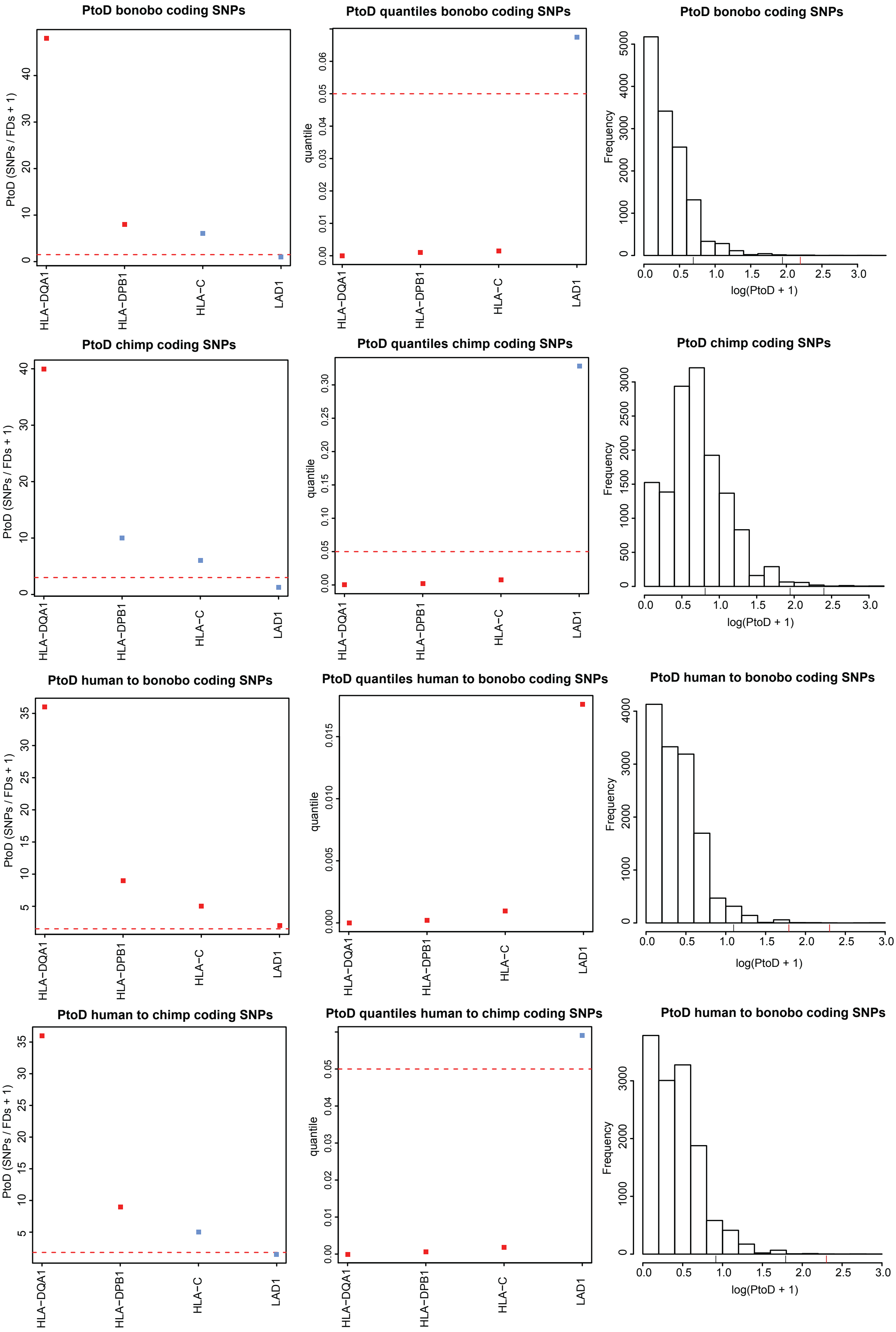
- 'coding+non-coding' (the entire set of SNPs found in the gene),
- 'coding' (only coding SNPs found in the gene) and
- '500bp' (all SNPs found in the +/-250bp window surrounding a trSNP). In this case, and for genes that have more than one shSNP, the *PtoD* value represent the average of *PtoD* values obtained for the 500bp windows around each of the

trSNPs. So assuming one single SNP is under selection, this is likely an underestimate of the diversity in the 500bp region around the trSNP maintained by long-term balancing selection.

PtoD values were calculated separately for each of the three species. To calculate polymorphism (P) we considered the number of SNPs found in each species. To calculate divergence (D, the number of fixed differences) we proceeded as follows: i) for bonobo and chimpanzee, we used the number of SNSs relative to human; ii) for human, we performed two separate comparisons using the number of SNSs relative to bonobo and to chimpanzee, separately. The results are shown in figures S5 and S6, and in tables S1 and S2.

In all comparisons, the candidates' group was significantly more diverse than the control group (all *P* < 2.05e-03 – see figure S3 and table S2). In all species, and considering the three different comparisons, *HLA-DQA1* showed the highest *PtoD* values in all species and for all sets of comparisons, with *LAD1* being the least polymorphic of the group. Looking at the different species, chimpanzee showed the less – but still highly – significant increase in diversity for the candidates' group (3.93e-04 < *P* < 2.05e-03) with 3 genes with P<0.05 (most likely due to the higher effective population size of the central chimpanzees compared to human and bonobo), followed by human (2.98e-04 < *P* < 4.25e-04) with 3-4 genes with P<0.05, and bonobo (3.05e-04 < *P* < 4.84e-04) with 3-4 genes with P<0.05 (table S2). Details for each species, set and gene are shown in table S1.


PtoD bonobo ALL SNPs
PtoD (SNPs / FDs + 1)
10
20
30
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles bonobo ALL SNPs
quantile
0.000
0.005
0.010
0.015
0.020
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD bonobo ALL SNPs
Frequency
0
1000
2000
3000
4000
5000
6000
0.0
0.5
1.0
1.5
2.0
2.5
PtoD chimp ALL SNPs
PtoD (SNPs / FDs + 1)
10
20
30
40
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles chimp ALL SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
0.06
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD chimp ALL SNPs
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)
PtoD human to bonobo ALL SNPs
PtoD (SNPs / FDs + 1)
0
10
20
30
40
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles human to bonobo ALL SNPs
quantile
0.000
0.005
0.010
0.015
0.020
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD human to bonobo ALL SNPs
Frequency
0
1000
2000
3000
4000
5000
0.0
0.5
1.0
1.5
2.0
2.5
log(PtoD + 1)
PtoD human to chimp ALL SNPs
PtoD (SNPs / FDs + 1)
0
10
20
30
40
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles human to chimp ALL SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
0.06
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD human to chimp ALL SNPs
Frequency
0
1000
2000
3000
4000
5000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)


PtoD bonobo coding SNPs
PtoD (SNPs / FDs + 1)
0
10
20
30
40
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD quantiles bonobo coding SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
0.06
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD bonobo coding SNPs
Frequency
0
1000
2000
3000
4000
5000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)
PtoD chimp coding SNPs
PtoD (SNPs / FDs + 1)
0
10
20
30
40
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD quantiles chimp coding SNPs
quantile
0.00
0.05
0.10
0.15
0.20
0.25
0.30
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD chimp coding SNPs
Frequency
0
500
1000
1500
2000
2500
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)
PtoD human to bonobo coding SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
25
30
35
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD quantiles human to bonobo coding SNPs
quantile
0.000
0.005
0.010
0.015
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD human to bonobo coding SNPs
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)
PtoD human to chimp coding SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
25
30
35
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD quantiles human to chimp coding SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
HLA−DQA1
HLA−DPB1
HLA−C
LAD1
PtoD human to bonobo coding SNPs
Frequency
0
1000
2000
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)

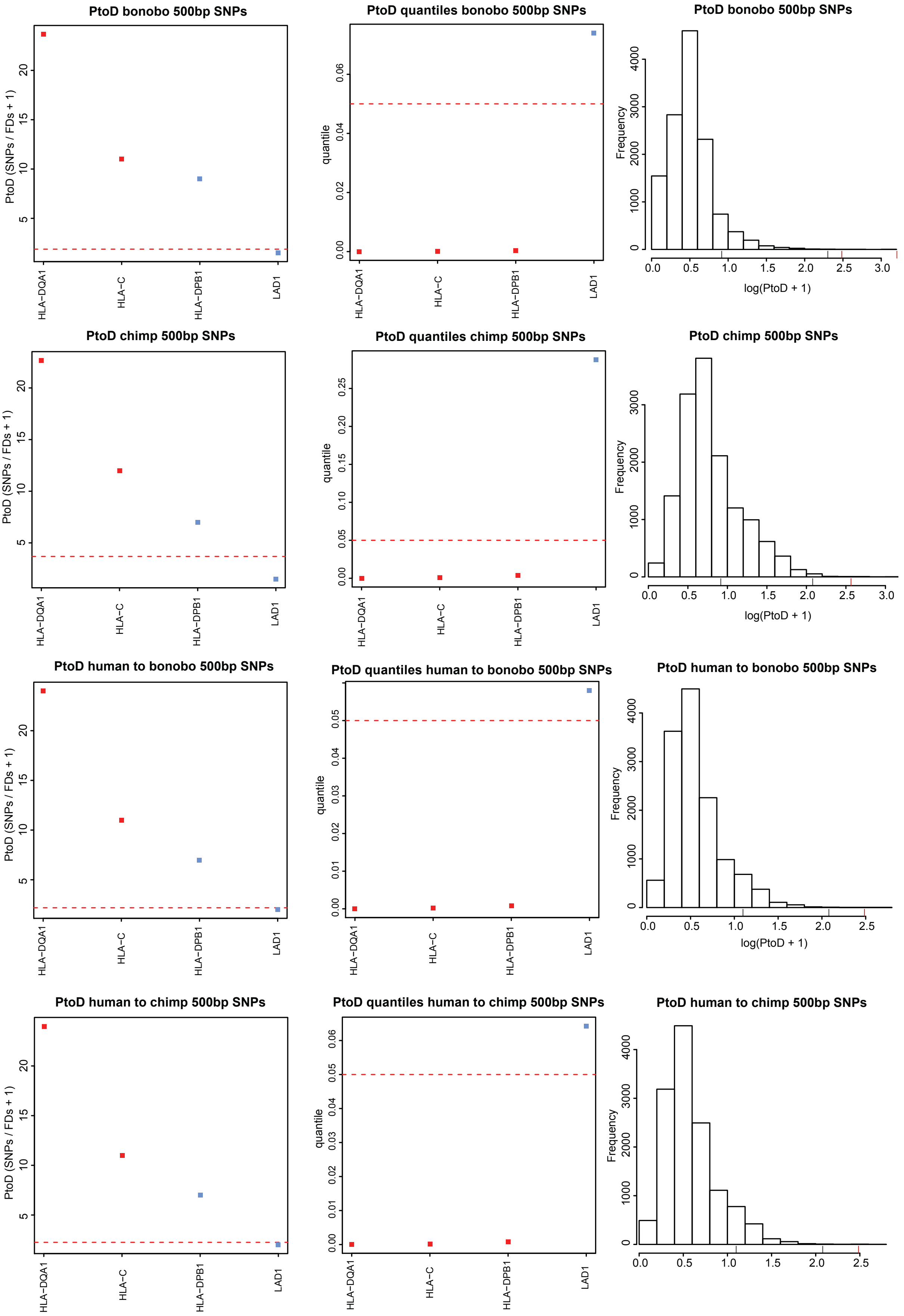

PtoD bonobo 500bp SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles bonobo 500bp SNPs
quantile
0.00
0.02
0.04
0.06
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD bonobo 500bp SNPs
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)
PtoD chimp 500bp SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles chimp 500bp SNPs
quantile
0.00
0.05
0.10
0.15
0.20
0.25
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD chimp 500bp SNPs
Frequency
0
1000
2000
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
log(PtoD + 1)
PtoD human to bonobo 500bp SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles human to bonobo 500bp SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD human to bonobo 500bp SNPs
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5
log(PtoD + 1)
PtoD human to chimp 500bp SNPs
PtoD (SNPs / FDs + 1)
5
10
15
20
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD quantiles human to chimp 500bp SNPs
quantile
0.00
0.01
0.02
0.03
0.04
0.05
0.06
HLA−DQA1
HLA−C
HLA−DPB1
LAD1
PtoD human to chimp 500bp SNPs
Frequency
0
1000
2000
3000
4000
0.0
0.5
1.0
1.5
2.0
2.5

Figure S5: PtoD ratios in candidate genes considering different sets of SNPs: ‘ALL’ includes all SNPs and FDs found in each gene; ‘coding’ represents variants found in the exons; and ‘500bp’ represents the average PtoD values for 500bp (-/+ 250bp) windows surrounding shSNPs in each gene. Left plots show the actual *PtoD* ratios for the candidate genes in each species and for each set, separately. The quantile values for each gene (considering all targeted genes) are shown in the middle, whereas its distribution can be seen in the histogram on the right (red bars represent genes for which P<0.05).

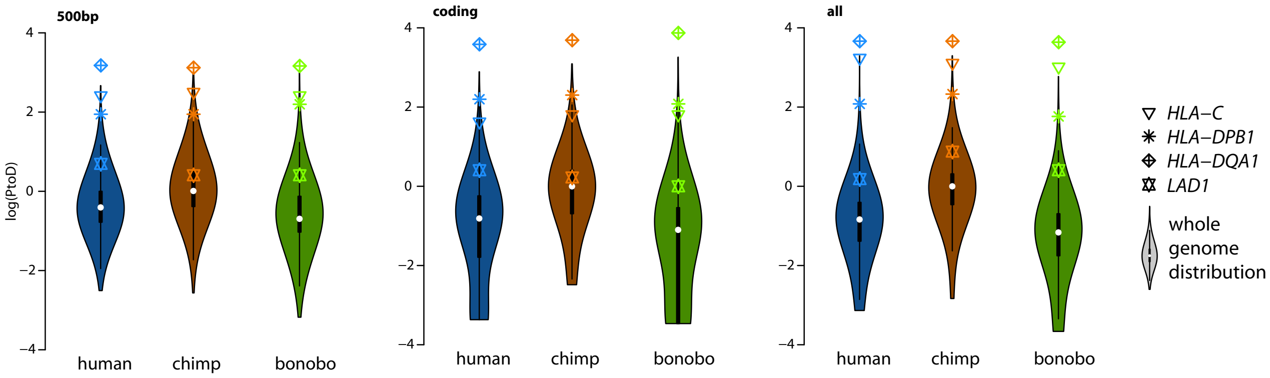


Figure S6: Violin plots of PtoD distributions of the controls (dark color) and the four genes (light color with symbols specified in the legend). The values are calculated: in 500 bp window centered on the SNP **(A)**; using only coding exonic regions **(B)**; and for the complete genes’ sequence **(C)**. The plots were created using the R function ‘vioplot’ from ‘vioplot’ package (Hintze, Nelson 1998) with default parameters.

PtoD ALL													
rank	**Gene**	***bonobo***			***chimp***			***human to bonobo***			***human to chimpanzee***		
		PtoD	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***
1	*HLA-DQA1*	38.00	1.55E-03	0.000	39.00	2.01E-03	0.000	39.00	1.60E-03	0.000	39.00	2.02E-03	0.000
2	*HLA-C*	20.00	8.30E-03	0.000	22.00	1.08E-02	0.000	25.00	8.54E-03	0.000	25.00	1.08E-02	0.000
3	*HLA-DPB1*	5.83	4.86E-02	0.001	10.25	6.38E-02	0.002	5.33	5.03E-02	0.000	8.00	6.47E-02	0.001
4	*LAD1*	1.50	NA	0.019	2.40	NA	0.059	1.50	NA	0.023	1.20	NA	0.061

PtoD coding													
rank	**Gene**	***bonobo***			***chimp***			***human to bonobo***			***human to chimpanzee***		
		PtoD	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***
1	*HLA-DQA1*	48.00	2.55E-03	0.000	40.00	1.04E-02	0.000	36.00	1.55E-03	0.000	36.00	2.01E-03	0.000
2	*HLA-C*	8.00	1.40E-02	0.001	10.00	5.38E-02	0.002	9.00	8.37E-03	0.000	9.00	1.09E-02	0.001
3	*HLA-DPB1*	6.00	8.61E-02	0.002	6.00	2.83E-01	0.008	5.00	4.99E-02	0.001	5.00	6.57E-02	0.002
4	*LAD1*	1.00	NA	0.067	1.25	NA	0.328	2.00	NA	0.018	1.50	NA	0.059

PtoD 500bp													
rank	**Gene**	***bonobo***			***chimp***			***human to bonobo***			***human to chimpanzee***		
		PtoD	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***	**PtoD**	**MW-U p-value**	***P***
1	*HLA-DQA1*	23.67	2.23E-03	0.000	22.67	8.03E-03	0.000	24.00	2.08E-03	0.000	24.00	2.17E-03	0.000
2	*HLA-C*	11.00	1.20E-02	0.000	12.00	4.22E-02	0.001	11.00	1.11E-02	0.000	11.00	1.17E-02	0.000
3	*HLA-DPB1*	9.00	7.28E-02	0.000	7.00	2.35E-01	0.003	7.00	6.65E-02	0.001	7.00	6.97E-02	0.001
4	*LAD1*	1.50	NA	0.074	1.50	NA	0.288	2.00	NA	0.058	2.00	NA	0.064

Table S1: PtoD values for the set of candidate genes. *MW-U P* is the recalculated Mann-Whitney U p-value (when comparing the candidate and the control sets) after removing the top-score gene from the candidate set. *P* is the percentile of each gene in the overall distribution.

PtoD			# genes	
set		MW-U *P*	MW-U < 0.05	P <0.05
bonobo	**ALL SNPs**	3.05E-04	4	4
	coding SNPs	4.84E-04	2	3
	500bp	4.34E-04	2	3
chimpanzee	**ALL SNPs**	3.93E-04	2	3
	coding SNPs	2.05E-03	1	3
	500bp	1.59E-03	2	3
human to bonobo	**ALL SNPs**	3.14E-04	2	4
	coding SNPs	2.98E-04	4	4
	500bp	4.07E-04	2	3
human to chimpanzee	**ALL SNPs**	3.95E-04	2	3
	coding SNPs	3.88E-04	2	3
	500bp	4.25E-04	2	3

Table S2: MW-U p-values comparing PtoD in candidate and control genes (genomic distribution of diversity). The numbers of candidate genes that are significant in the MW-U ranked test, as well as the number of genes in the top 5% of the distribution are also shown.

V – Tajima’s D

A classical test to detect departures from neutrality in the genome is Tajima’s D (Tajima 1989). Particularly, a positive value of Tajima’s D – caused by an excess of intermediate-frequency alleles – is classically considered as a signature of balancing selection in the absence of population substructure. We calculated Tajima’s D for the set of trSNP-containing genes and a set of control genes (considering all genes with at least six polymorphic sites, which is the minimum number of SNPs in the four candidate genes in all species). Because we targeted the exons, Tajima’s D for the control set are expected to have on average slightly negative values due to the action of purifying selection, and we observe that shift (Figure S9). On the contrary, all candidate genes have positive Tajima’s D, with the exception of *LAD1* in bonobo. These values, however, are not all significantly higher when compared with the control genes (5% upper tail cutoff, Figure S9). We note that the power of this test is hampered by the limited number of SNPs in the coding regions of genes.

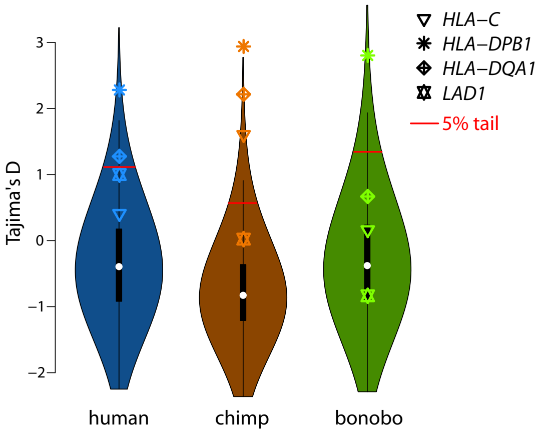


Figure S9 – Violin plots representing the distribution of Tajima’s D in control genes (dark color) and candidate genes (represented by symbols) in all species. The plots have been generated with the ‘vioplot’ function in R (Hintze, Nelson 1998) using default values. The red line represents the 5% upper tail boundary of the distribution for each species.

VI – Comparison with available datasets

In order to compare our results to previously published studies, we investigated whether additional shSNPs in candidate genes (that we might have missed) were present in a whole-genome dataset consisting of several sequenced individuals from different great ape species (Prado-Martinez et al. 2013). We also verified whether some of the trSNPs uncovered in our study were found in a genome-wide scan for long-term balancing selection in humans and chimpanzees (Leffler et al. 2013).

If we compare the trSNPs found in this study to the dataset of Prado-Martinez et al. (2013), out of the 8 trSNPs uncovered in our study, 6 are also shared between the three species in that dataset, including rs12088790. Moreover, and now considering SNPs shared among pairs of species in *LAD1*, we identified two additional intronic shSNPs, one shared between human and bonobo (chr1: 201348361), and the other shared between human and orangutan (chr1: 201361157). We have not investigated whether these two shSNPs show signatures of balancing selection, as they are not shared between the three species analyzed in this work and because they are located far away from rs12088790 (>5kb). Apart from these, we also found several cSNPs in *LAD1*, shared across different species, although they are associated with CpG sites and, therefore, are likely the result of recurrent mutations (data not shown).

However, the picture looks different when we attempt to retrieve our eight trSNP from the set of human-chimpanzee trSNPs that were identified as part of short trans-species haplotypes in Leffler et al. (2013), as we can detect none. This is probably due to the different strategies adopted in the studies, as we focused our analysis on shared polymorphism on the coding sequences of the genome, while Leffler et al. (2013) focused on shared haplotypes (with at least 2 SNPs in perfect linkage), which happened to be largely non-coding. We though also searched for the presence of our trSNPs in a list of single coding shSNPs provided by Leffler et al. (2013), and retrieved none. Although perhaps surprising, the lack of correspondence might be explained by a number of differences between the two studies regarding samples and coverage depth. For example, we analyze 20 individuals per species with an average coverage of ~18X in all species; Leffler et al. (2013) analyzed a genome-wide dataset with only moderate coverage (~9X for the chimpanzees and 3.4X for the human samples), with smaller chimpanzee sample size (10 individuals) and much larger human sample sizes (59 individuals) than our dataset. In addition, the two studies analyzed different chimpanzee subspecies (*Pan troglodytes troglodytes* vs *Pan troglodytes verus*). Nevertheless, we note that Leffler et al. (2013) uncovered human-chimpanzee shSNPs in the two HLA genes where we identify trSNPs, only the SNPs identified are different (4 shSNPs in *HLA-DQA*, 3 shSNPs in *HLA-DPB1*.

VII – Supplementary Tables

human	chimpanzee	bonobo
NA18501	Agnagui	Api
NA18504	Bailele	Bandundu
NA18505	Bayokele	BilliL
NA18508	Bimangou	Boende
NA18516	Botsomi	BoloboGelcut
NA18522	Casimir	Fizi
NA18523	CastroGelcut	Isiro
NA18853	Chinoc	Keza
NA18856	ClaraT	Kikwit
NA18858	Dzeke	Kisantu
NA18861	Elikia	Kubulu
NA18870	FanTuek	Likasi
NA18871	Gao	Lipopo
NA18912	Golfi	Lodja
NA19093	GrandMaitre	Lomami
NA19102	Imphondo	MalouL
NA19137	Loufoumbou	Matadi
NA19138	Lufino	Max
NA19238	Marcelle	Semendwa
NA19239	Moka	Tshilomba

Table S3: The 20 humans, 20 chimpanzees and 20 bonobos used in this study.

POP	P
ASW	0.0008
LWK	0.0019
YRI	0.0262
CEU	0.9999
FIN	0.4318
GBR	0.9763
TSI	0.9768
CHB	0.3338
CHS	0.7786
JPT	0.2588
MXL	0.9632
CLM	0.8802
PUR	0.1560

Table S4: P-values of the MWU test between the MAF of *LAD1* and that of the entire chromosome 1 in the 1000Genomes (Abecasis et al. 2012) populations. Values closer to 0 and 1 indicate shift towards intermediate- and low-frequency alleles, respectively. We filtered the 1000Genomes data by considering only SNPs that: 1) are in the 50mer mappability track (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability); 2) are not in the Tandem Repeat Finder; 3) are not in annotated segmental duplications (Cheng et al. 2005; Alkan et al. 2009; Prufer et al. 2012); and 4) are perfectly aligned to PanTro2 genome.

Mutations	human		chimpanzee		bonobo	
	all	shGenes	all	shGenes	all	shGenes
Synonymous (S)	18,955	21	43,023	21	15,549	19
Non-Synonymous (NS)	18,208	26	36,151	31	15,079	30
NS/S ratio	0.96	1.23	0.84	1.48	0.97	1.44

Table S5: Number of synonymous and non-synonymous SNPs for each species using all coding SNPs and those falling within the four candidate genes ‘shGenes’. χ^2 test of differences between ‘all’ and ‘shGenes’ for each species are all not significant.

References

Abecasis, GR, A Auton, LD Brooks, MA DePristo, RM Durbin, RE Handsaker, HM Kang, GT Marth, GA McVean. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65.

Alkan, C, JM Kidd, T Marques-Bonet, et al. 2009. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet 41:1061-1067.

Andres, AM. 2011. Balancing Selection in the Human Genome. Encyclopedia of Life Sciences (eLS). Chichester: John Wiley & Sons Ltd.

Asthana, S, S Schmidt, S Sunyaev. 2005. A limited role for balancing selection. Trends Genet 21:30-32.

Benzer, S. 1961. On the Topography of the Genetic Fine Structure. Proc Natl Acad Sci U S A 47:403-415.

Charlesworth, D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64.

Cheng, Z, M Ventura, X She, et al. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88-93.

Clark, AG. 1997. Neutral behavior of shared polymorphism. Proc Natl Acad Sci U S A 94:7730-7734.

Coulondre, C, JH Miller, PJ Farabaugh, W Gilbert. 1978. Molecular basis of base substitution hotspots in Escherichia coli. Nature 274:775-780.

Cutrera, AP, EA Lacey. 2007. Trans-species polymorphism and evidence of selection on class II MHC loci in tuco-tucos (Rodentia: Ctenomyidae). Immunogenetics 59:937-948.

Graser, R, C O'HUigin, V Vincek, A Meyer, J Klein. 1996. Trans-species polymorphism of class II Mhc loci in danio fishes. Immunogenetics 44:36-48.

Griffiths, RC, S Tavare. 1994. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci 344:403-410.

Hintze, JL, RD Nelson. 1998. Violin plots: a box plot-density trace synergism. The American Statistician 52(2):181-184.

Hodgkinson, A, A Eyre-Walker. 2010. The genomic distribution and local context of coincident SNPs in human and chimpanzee. Genome Biol Evol 2:547-557.

Hodgkinson, A, A Eyre-Walker. 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12:756-766.

Hodgkinson, A, E Ladoukakis, A Eyre-Walker. 2009. Cryptic variation in the human mutation rate. PLoS Biol 7:e1000027.

Hudson, RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337-338.

Johnson, PL, I Hellmann. 2011. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. Genome Biol Evol 3:842-850.

Kikkawa, EF, TT Tsuda, D Sumiyama, et al. 2009. Trans-species polymorphism of the Mhc class II DRB-like gene in banded penguins (genus Spheniscus). Immunogenetics 61:341-352.

Klein, J, Y Satta, C O'HUigin, N Takahata. 1993. The molecular descent of the major histocompatibility complex. Annu Rev Immunol 11:269-295.

Leffler, EM, Z Gao, S Pfeifer, et al. 2013. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science 339:1578-1582.

Loisel, DA, MV Rockman, GA Wray, J Altmann, SC Alberts. 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 5' cis-regulatory region. Proc Natl Acad Sci U S A 103:16331-16336.

McKenna, A, M Hanna, E Banks, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-1303.

Prado-Martinez, J, PH Sudmant, JM Kidd, et al. 2013. Great ape genetic diversity and population history. Nature 499:471-475.

Prufer, K, K Munch, I Hellmann, et al. 2012. The bonobo genome compared with the chimpanzee and human genomes. Nature 486:527-531.

Rozen, S, H Skaletsky. 2000. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132:365-386.

Segurel, L, EE Thompson, T Flutre, et al. 2012. The ABO blood group is a trans-species polymorphism in primates. Proc Natl Acad Sci U S A 109:18493-18498.

Sutton, JT, BC Robertson, CE Grueber, JA Stanton, IG Jamieson. 2013. Characterization of MHC class II B polymorphism in bottlenecked New Zealand saddlebacks reveals low levels of genetic diversity. Immunogenetics 65:619-633.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Tavare, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor Popul Biol 26:119-164.

Wiuf, C, K Zhao, H Innan, M Nordborg. 2004. The probability and chromosomal extent of trans-specific polymorphism. Genetics 168:2363-2372.